International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:6, No:8, 2012

Data Mining Classification Methods Applied in Drug Design

Mária Stachová,Lukáš Sobíšek

Abstract**—**Data mining incorporates a group of statistical methods used to analyze a set of information, or a data set. It operates with models and algorithms, which are powerful tools with the great potential. They can help people to understand the patterns in certain chunk of information so it is obvious that the data mining tools have a wide area of applications. For example in the theoretical chemistry data mining tools can be used to predict moleculeproperties or improve computer-assisted drug design. Classification analysis is one of the major data mining methodologies. The aim of thecontribution is to create a classification model, which would be able to deal with a huge data set with high accuracy. For this purpose logistic regression, Bayesian logistic regression and random forest models were built using R software. TheBayesian logistic regression in Latent GOLD software was created as well. These classification methods belong to supervised learning methods.

It was necessary to reduce data matrix dimension before construct models and thus the factor analysis (FA) was used. Those models were applied to predict the biological activity of molecules, potential new drug candidates.

Keywords**—**data mining, classification, drug design, QSAR

I. Introduction

THE area of drug design can be considered a very attractive from many points of view. The main point is definitely the growth of civilization diseases such as cancer.The primarytask of this paper is to present statistical classification models- or structure activity relationship (SAR) models that can be used to predict the biological activity of a molecule (potential new drug candidate) in high accuracy.

The biological activity of molecules is usually measured in assays to establish the level of inhibition of particular signal transduction or metabolic pathways. Chemicals can also be biologically active by being toxic. Drug discovery often involves the use of QSAR (Quantitative Structure-Activity Relationship), or simple SAR (Structure-Activity Relationship), to identify chemical structures that could have good inhibitory effects on specific targets and have low toxicity (non-specific activity). For example: activation of Vascular Endothelial Growth Factor Receptor-2 (VEGFR-2) is essential for new vessel initiation in early stages of angiogenesis through induction of proliferation, migration, and survival of endothelial cells. Inhibition of VEGFR-2 activity may therefore become an efficient way to stabilize or slow down the progression of solid tumors in the anti-angiogenesis therapy[1].

M. Stachováiswiththe Department ofQuantitativeMethods and InformationSystems, FacultyofEconomics, Matej BelUniversity, Banská Bystrica, SK 97590 Slovak Republic (e-mail: maria.stachova@umb.sk).

L. Sobíšekispostgraduate student oftheDepartmentofStatistics and Probability, FacultyofInformatics and Statistics, UniversityofEconomics, Prague, CZ13067CzechRepublic (e-mail: lukas.sobisek@yahoo.com).

In this contribution we use data mining methods to improve computer-assisted drug design. We created four classification models based on different statistical approaches that can help us to select just a few molecules among hundreds of thousands. Classification analysis belongs to supervised learning methods. The models help us to identify molecules, which might be classified as biological active and therefore it is worth involving them intothe laboratory tests. The rate of molecular activity is IC50 factor. This parameter describes how effective the drug is. It indicates how much of a particular drug (molecule in our case) is neededto inhibit a given biological process by half. Lower value indicates greater potency.

This type of analysis helps to optimize the search for relevant new drug structures and therefore has major importance for the industry.

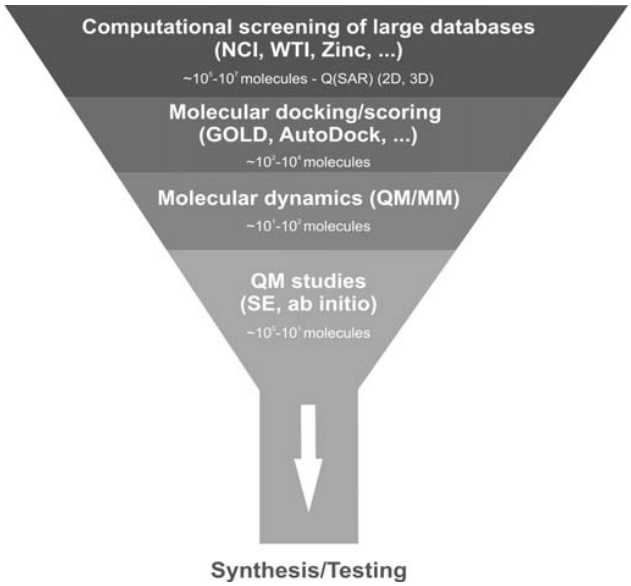

Fig. 1 Process of new medications searching

As the Fig. 1 illustrates, our analysis belongs to the first stage of theprocess of finding new medications and it plays an important role in reducing the large data into a useable small subset. The next stages are the molecular scoring and docking, the methods of molecular dynamics (quantum mechanical or molecular mechanical) and the quantum mechanical semi-empirical or ab-initio studies. The process ends with the synthesis and the laboratory testing of new drugs candidates.

II. Data Description

A database of 1252 molecules with known VEGFR-2 inhibition activity (IC50) values, collected by AndrejBoháč[2], is considered as a training group. This data set is not publicly available and it was obtainedin a private

communication.Based on experimental knowledge we set a border line of IC50 level equal to 50. Each of the molecules with less IC50 was coded as a potential drug candidate and per contra a molecule with higher IC50 was coded as a non drug candidate. A set of 1818 characteristics, so called molecular descriptors, for each of these molecules was computed by Dragon software[3]. The obtained data were used as a training data to build classification models. Constructed models were used to select molecules, potential new drugs, from the data set, without known IC50 values. The new data set involved 525 270 molecules was obtained from the public database ZINC [4]. The set of the same Dragondescriptors was computed for molecules included in this database.

III. Methodology

The first step of our analysis was to reduce the high data dimensionality. For this purpose we used the factor analysis which models *k*-latent variables (factors) F_i, $i = 1, 2,\ldots k$ explaining the variations and covariations of *p*-variables at the best. Härdle and Simar in[5]describes this method in more detail. We inferred factors by IBM SPSS Statistics 18 software. We came out of the correlation matrix and chose principle component method without rotation for factor extraction. We obtained 142 factors. Factor score coefficients were calculated by Bartlett method [6]. We used factors as new predictors.

The second step was to build different classification models on our training data set as a regression model implemented in Latent GOLD software, Bayesian regression model, logistic regression and random forest models all constructed in R software.

Latent Class (LC) Regression model implemented in Latent GOLD software is a model with a single nominal latent variable x with K-classes. Each class represents a homogeneous group of instances having identical regression coefficients. Latent GOLD offers several types of regression models [7]. We chose binary logistic regression model on account of dichotomous dependent variable (IC50). LC Regression model is based on the classical finite mixture model, which assumes conditional independence of variables.
Probability structure of the regression model [8]takes the following formula:

$$f\left(Y_i\left|Z_i\right.\right)=\sum_{x=1}^{K}P\left(x\left|Z_i^{cov}\right.\right)f\left(\mathbf{y}_i\left|x,Z_{it}^{pred}\right.\right) \quad (1)$$
$$=\sum_{x=1}^{K}P\left(x\left|Z_i^{cov}\right.\right)\prod_{t=1}^{T_i}f\left(y_{it}\left|x,Z_{it}^{pred}\right.\right).$$

The value of dependent variable for instance i at replication t is denoted by y_{it}, and its total number of replications by T_i. A vector of exogenous variables for $i$is depicted Z_i.Q predictorsz_{itq}^{pred} affect y_{it}.R covariates z_{ir}^{cov}affecting$x$.

Our other models were obtained using statistical system R.In addition to the basic R package we used 'randomForest'[9]and 'arm'[10]packages.

The 'randomForest' package was used to construct a random forest model. This model is a machine learning ensemble classifier that consists of many (500 in our case) decision trees. The computational procedure of decision trees is well described in [11], [12].

The algorithm for inducing a random forest was developed by Leo Breiman[13]and Adele Cutler, and "Random Forests" is their trademark. The term came from random decision forests and it was first proposed by TinKam Ho of Bell Labs in 1995 [14]. The method combines Breiman's "bagging" idea and Ho's "random subspace method" to construct a collection of decision trees with controlled variations. The variance reduction is achieved in the tree-growing process through random selection on a bootstrapped dataset.In case of classification random forest, the number of $\sqrt{p}$ as input variables from p variables is selected as candidates for splitting. Let assume that $\{T_b\}_1^B$ is ensemble of trees and θ_b characterizes the b^{th} random forest tree in terms of split variables, cut points at each node, and terminal-node values than after B such trees $\{T(x;\theta_b)\}_1^B$ are grown, the random forest predictor is

$$C_{rf}^{B}=majority\ _vote\left\{\hat{C}_b(x)\right\}_1^B. \quad (2)$$

$\hat{C}_b(x)$is the class prediction of the b^{th} random forest tree. From (2) is obvious that if we want to classify a new entry by random forest, we let it go through each decision tree. Each tree has a vote and we say that the trees are voting for the most popular class.

Reference[15]is more detailed in the classification random forest algorithm.

The next model we built using R was the well known logistic regression model:

$$f\left(Y_i\left|X_i\right.\right)=\text{logit}^{-1}\left(\beta.X_i\right)=\frac{1}{1+e^{-\beta.X_i}}, \quad (3)$$

where Y_i is an outcome, X_i is a vector of explanatory variables and β is a vector of regression coefficient.

We used *glm*() function in R with threshold set at point 0.5.

An algorithm of this model can be found in [15].

To build the last model, Bayesian regression model, we use *bayesglm*() function that is a part of 'arm' package in R software. It is a Bayesian function for generalized linear modeling. It is a simple alteration of *glm*() function that uses an approximate expectation maximization (EM) algorithm to update the logistic regression coefficients at each step using an augmented regression to represent the prior information.Let's assume that coefficients $\beta_j$have independent t prior distributions with centers μ_j and scales s_j. The idea of EM approximation is to express the t prior distribution for each coefficient β_j as a mixture of normals with unknown scale σ_j:

$$\beta_j \sim N(\mu_j, \sigma_j^2), \quad \sigma_j^2 \sim \text{Inv-}\chi^2(\nu_j, s_j^2) \tag{4}$$

and then average over the β_j 's at each step, treating them as missing data and performing the EM algorithm to estimate the σ_j's. The algorithm continues with alternating of iteratively weighted least squares and of EM in particular steps. After reaching approximate convergence, we get an estimate and covariance matrix for the vector β and the estimated σ_j 's.

An algorithm of this approach is described in detail in [16].

The complete R-code usedin the paper is available upon request.

IV. Results

Quality of a certain model is often described by confusion matrix. In this matrix each row represents the instances in an actual class, while each column represents the instances in a predicted class. We applied a Bayesian logistic regression constructed in R and Latent GOLD software on our training data and we achieved the matrices shown in Table I and Table II. The error rate of the first model was 16.6%.

TABLE I
Bayesian Logistic Regression In R

	FALSE	TRUE
FALSE	807	77
TRUE	131	237
Error rate		16.60%

The second model was unsuccessful in 36.6% of cases.

TABLE II
Bayesian Logistic Regression In Latent Gold

	FALSE	TRUE
FALSE	620	21
TRUE	437	174
Error rate		36.60%

Table III illustrates how the logistic regression was successful in classification. Error rate of this model is 16.8%.

TABLE III
Logistic Regression In R

	FALSE	TRUE
FALSE	805	79
TRUE	131	237
Error rate		16.80%

Table IV shows the success of the random forest model built in R software. This model failed in 22% of cases.

TABLE IV
Random Forest In R

	FALSE	TRUE
FALSE	851	33
TRUE	242	126
Error rate		22.00%

After comparing the efficacy of these four models we chose three of them to be applied on our test data set including 525 270 molecules. We excluded the regression model from Latent Gold software because of a lack of his predictive capability. We also made this decision because the software was not able to allocate enough internal memory from available RAM and thus it was not efficientto process such a large data. This problem did not occur with R software, because after adding the option to increase the available memory, the 64-version of this software was able to use almost 20 GB of RAM. All our analyses were performed on a personal computer with 24 GB RAM and 3.07 GHz processor.

Table V shows that the random forest model selected 10,124 molecules from test data set and stated them as biological active. Logistic regression selected 618 molecules and logistic regression with Bayesian approach selected 917 molecules as possible drugs. Our analysis continued by combining these results. After overlapping the data we found 19 molecules mutual for all selection.

TABLE V
Results From Models Selections With Overlap

	Number of selected molecules	overlap
Bayesian logistic regression	917	
Logistic regression	618	19
Random forest	10124	

This set of 19 molecules was subsequently sent to the laboratory of theoretical chemistry at Matej Bel University to be scored from the molecular modeling point of view. The results showed that our molecules can be considered promising antiangiogenic agents. Moreover these molecules fulfill theoretical criteria to be new drug candidates.

V. Conclusion

Usage of classification methods is effective. Their predictive ability is sufficient. Logistic regression (both Bayesian and classical) assigns molecules with more precise accuracy then other methods.Therefore these models can be applied for prediction of new molecules, potential drugs candidates. Combination of different classifiers is also very useful. It helps us to reduce the large data into useable small subset.

Acknowledgment

The research was supported by the Grant Agency of Matej Bel University UGA, project I-10-005-08/101610 and by VSE - PragueprojectIGA VSE F4/6/2012.

We thank Andrej Boháč for providing us the training set of molecules and we also thank Marek Skoršepa for scoring our selected molecules.

References

[1] A.Hoeben, B.Landuyt, M. S. Highley, H.Wildiers, A. T. Van Oosterom, and E. A. De Bruijn,"Vascular Endothelial Growth Factor and Angiogenesis," *Pharmacological Reviews*, vol. 56 no. 4, pp. 549-580, Dec. 2004.

[2] Boháč A., Faculty of Natural Science, Comenius University in Bratislava, andrej.bohac@fns.uniba.sk, private communication, 2009.

[3] DRAGON Professional verzion 5.5 2007,TALETE, srl.
[4] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC--a free database of commercially available compounds for virtual screening,"*J. Chem. Inf. Model*., 2012, accepted for publication.
[5] W. Härdle, and L. Simar,*Applied Multivariate Statistical Analysis*.New York: Springer, Berlin, 2007.
[6] IBM SPSS Statistics, Help, Algorithms [online]. On-line manual. [cit. 2011-01-16], 127.0.0.1:4004/help/index.jsp?topic=/com.ibm.spss.statistics.help/alg_introduction.htm.
[7] J. K. Vermunt, and J. Magidson,*Technical Guide for Latent GOLD 4.0: Basic and Advanced* [online]. Statistical Innovations Inc., Belmont Massachusetts, 2005, [cit. 2011-01-16]. www.statisticalinnovations.com/products/LGtechnical.pdf.
[8] J. K.Vermunt, and J. Magidson,"Latent class cluster analysis,"*J. A*. Hagenaars, A. L. McCutcheon (eds.). *Applied Latent Class Analysis*. Cambridge : Cambridge University Press, pp. 89-106,2002.
[9] A. Liaw, and M. Wiener,"Classification and Regression by randomForest," *R News*, vol. 2, no.3, pp. 18—22,2002.
[10] A.Gelman, Y. S. Su, M.Yajima, J. Hill, M. G.Pittau, J. Kerman, and T. Zheng, "arm: Data Analysis Using Regression and Multilevel/Hierarchical Models," R package version 1.5-02.://CRAN.R-project.org/package=arm,2012.
[11] L. Breiman, J.H. Friedman, R.A. Olshen, andC.J. Stone,*Classification and Regression Trees*.Chapman and Hall,Wadsworth, Inc., New York, 1984.
[12] StatSoft, Inc. Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/, 2011.
[13] L. Breiman, " Random Forests,"in *Machine Learning*, vol. 45, pp. 5-32, 2001.
[14] Ho Tin Kam,"Random Decision Forest," in.*Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition*, Montreal, Canada, August 14-18, pp. 278-282, 1995.
[15] T. Hastie, R.Tibshirani, and J. H. Friedman,*The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag, 2001.
[16] A.Gelman, A.Jakulin, M. G.Pittau, and Y.S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The annals of Applied Statistics*, vol. 2, no. 4, pp.1360-1383, 2008.